RobTrack: A Robust Tracker Baseline towards Real-World Robustness in Multi-Object Tracking and Segmentation

Dongxu Wei[1,2], Jiashen Hua[2], Hualiang Wang[1,2], Baisheng Lai[2],
Kejie Huang[1], Chang Zhou[2], Jianqiang Huang[2], and Xiansheng Hua[2]
[1]Zhejiang University, [2]Alibaba Group

Abstract

Multi-Object Tracking and Segmentation (MOTS) aims to detect objects' masks and track their trajectories throughout videos, which has a wide range of applications in autonomous driving, safety monitoring, video analytics, etc. Most existing MOTS methods focus on tuning and competing over one specific dataset at a time, which can only reflect their performance on a small subset of real-world scenarios. However, real-world tracking requires trackers to work robustly and consistently on any dataset without dataset-specific tuning. To achieve this, we propose RobTrack, a robust tracker baseline optimized from two aspects w.r.t. robust detection and association. For robust detection, a strong Cascade Mask R-CNN detector is utilized to first generate coarse object masks, incorporated with a Massive Instance Augmentation (MIA) method during training to improve its generalization ability in various scenarios. Then, a Patch-Based Mask Refinement (PBMR) module is applied to these coarse object masks for higher mask quality. For robust association, we propose an Ensemble Full-Category (EFC) tracker consisting of two expert trackers w.r.t. common (e.g., person, vehicle) and general (e.g., non-human and non-vehicle objects) categories, enabling optimal association for both category domains. To further improve the association accuracy, a well-designed Offline Re-Linking (ORL) strategy is applied to each video sequence to obtain the final tracking results. Extensive experiments are conducted on a very large-scale RobMOTS dataset, which consists of eight smaller MOTS datasets. The results show that our proposed method performs well in both detection and association for all of the eight datasets, and achieves the first place in the RobMOTS 2021 challenge, demonstrating its significance in robust real-world tracking.

1. Introduction

Multi-Object Tracking and Segmentation (MOTS) aims at automatically detecting pixel-level object masks and consistently assigning their IDs throughout the video sequence. As demanded by the task of Robust MOTS (RobMOTS) [13], all the objects belonging to 80 COCO [11] categories in eight sub-benchmarks should be detected and tracked, which can cover most objects of interest in real-world scenarios. To encourage meaningful improvements for robust tracking, this task uses a robust HOTA [14] metric for evaluation, which treats the two sub-tasks of detection and association equally, and can balance the importance of class-averaged and detection-averaged tracking under real-world setting.

Since both detection and association matter, we propose to optimize our tracker from two aspects w.r.t. robust detection and association sub-tasks, respectively. For robust detection, we expect a detector to estimate high-quality masks for any object without biases to specific categories or scenarios, thus providing the subsequent association process with more accurate and robust object proposals. To achieve this, we propose to enhance our detector with Massive Instance Augmentation (MIA) during training and Patch-Based Mask Refinement (PBMR) during inference, aiming to improve the generalization ability (more robust) and mask quality (more accurate), respectively. For robust association, we expect a tracker to robustly track objects of all the categories. However, we conjecture that there exists a gap when tracking objects of different super-categories (i.e., common and general). To address this, we propose an Ensemble Full-Category (EFC) tracker that leverages expert knowledge from both of the two category domains to enable optimal association for both of them. Besides, to further improve the association accuracy, a simple Offline Re-Linking (ORL) method is applied to re-link broken trajectories belonging to the same IDs together. Experiments conducted on the RobMOTS dataset show our superiority to all of the other methods on the leader board for both detection and association accuracy, indicating its significance towards robust real-world tracking.

2. Method

The overview of our method is shown in Figure 1, which consists of two steps: robust detection and robust associa-

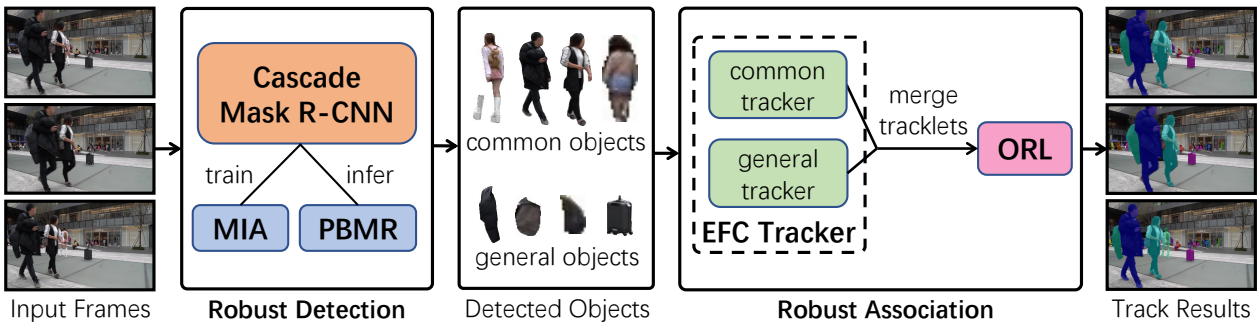

Figure 1. Overview of RobTrack.

tion, responsible for detecting objects and associating them throughout input frames, respectively.

2.1. Robust Detection

Our robust detector uses Cascade Mask R-CNN [2] with EfficientNet-B7 [24] as the backbone and NAS-FPN [7] (P3∼P7) as the feature pyramid. To improve its robustness, we further apply Massive Instance Augmentation (MIA) and Patch-Based Mask Refinement (PBMR) during training and inference detailed in the following.

2.1.1 Massive Instance Augmentation

Although COCO dataset [11] contains a large amount of training data for instance segmentation, it can only cover a small subset of various real-world scenarios. If we train our detector purely with the original COCO data, it probably can't generalize well to real-world scenes. One possible solution is to augment existing annotated samples to create massive new instance mask and image pairs. Besides of using traditional data augmentations like random resizing and flipping, we further utilize copy-paste augmentation presented in [6] to maximize the data efficiency. By pasting diverse objects of various scales to new background images, we can create challenging and novel training data that suits better for real-world setting, boosting the generalization ability and robustness of our detector.

2.1.2 Patch-Based Mask Refinement

Since the task of MOTS matches detected objects with ground truths using pixel-level Intersection-over-Union (IoU), mask quality can also affect the tracking performance. Therefore, we apply patch-based refinement [25] to previously detected coarse object masks to enhance our mask details. Specifically, we first extract a series of small image patches along the coarse mask boundary for each object. Then, a refinement network trained on COCO is applied to refine boundaries for these patches. After that, we reassemble these patches to formulate a new high-quality full mask.

2.2. Robust Association

Provided with robust and high-quality detection results, our robust association focuses on tracking the detected objects throughout videos. To achieve this, we propose an Ensemble Full-Category (EFC) tracker and an Offline Re-Linking strategy detailed below.

2.2.1 Ensemble Full-Category Tracker

For robust association, our tracker should be able to re-identify objects of all categories throughout videos. Experimentally, we observe that utilizing trackers based on different re-identification (reID) models for objects of different super-categories can achieve optimal association. To be more specific, we divide 80 COCO classes into two super-categories: common and general. "Common" refer to those that commonly seen in real-world scenarios, including persons and vehicles (i.e. car, bus, truck, motorcycle, bicycle). "General" refer to the opposite, including animals, accessories, still stuff and so on.

For "common", we let all the objects in a frame image share one multi-scale global feature, which is extracted by a ResNet-50 backbone with FPN neck [10]. Then, a shared reID head is added to the global feature to estimate object-specific reID features via RoI-Align [9]. Such feature sharing can provide each object with more contextual information such as location and background appearance, both are important reID clues in scenes with crowded people and vehicles, which is very common in real-world scenarios. We train this common reID model on RobMOTS dataset by filtering out non-person and non-vehicle annotations to focus on common objects. To enrich reID samples, we further incorporate a region proposal network into this model to enable quasi-dense reID sampling as described in [20]. With these common reID features, a completely appearance-based bi-directional matching [20] is utilized as the association strategy to track common objects, which we find is better than adding motion clues like optical flow [15, 12] and Kalman filter [1, 4, 5].

Table 1. Ablation studies on Massive Instance Augmentation (MIA), Patch-Based Mask Refinement (PBMR), Ensemble Full-Category (EFC) Tracker, and Offline Re-Linking (ORL) strategy. Here Tracker-C and Tracker-G denote single trackers with different reID implementations. HOTA-F, HOTA-C, HOTA-D denote final, class-averaged, detection-averaged HOTA scores, respectively. DetA, DetRe, DetPr denote detection accuracy, recall, precision, respectively. AssA, AssRe, AssPr denote association accuracy, recall, precision, respectively.

MIA	PBMR	Tracker	ORL	HOTA-F↑	HOTA-C↑	HOTA-D↑	DetA↑	DetRe↑	DetPr↑	AssA↑	AssRe↑	AssPr↑
-	-	EFC	-	60.51%	55.79%	65.95%	57.68%	64.45%	77.51%	65.21%	70.00%	82.95%
✓	-	EFC	-	62.32%	58.14%	67.03%	60.75%	68.10%	78.08%	65.53%	70.19%	84.15%
✓	✓	EFC	-	62.83%	58.31%	67.98%	61.32%	68.38%	79.17%	66.06%	70.54%	84.60%
✓	✓	Tracker-C	✓	62.28%	57.83%	67.34%	60.98%	68.10%	78.61%	65.36%	69.49%	84.79%
✓	✓	Tracker-G	✓	61.78%	57.43%	66.70%	60.95%	68.21%	78.43%	64.46%	67.79%	**85.64%**
✓	✓	EFC	✓	**63.15%**	**58.64%**	**68.29%**	**61.32%**	**68.38%**	**79.17%**	**66.72%**	**71.35%**	84.26%

Table 2. Comparisons with other methods on the leader board of RobMOTS challenge.

Method	split	HOTA-F↑	HOTA-C↑	HOTA-D↑	DetA↑	DetRe↑	DetPr↑	AssA↑	AssRe↑	AssPr↑
STP [13]	val	55.90%	51.43%	61.07%	57.62%	64.63%	77.11%	56.34%	59.67%	83.60%
SIA [23]	val	58.73%	54.22%	63.96%	57.64%	64.63%	77.15%	61.71%	65.18%	**85.28%**
SBT [26]	val	60.52%	56.25%	65.42%	57.68%	64.45%	77.51%	65.21%	70.00%	82.95%
Ours	val	**63.15%**	**58.64%**	**68.29%**	**61.32%**	**68.38%**	**79.17%**	**66.72%**	**71.35%**	84.26%
STP [13]	test	54.35%	48.88%	61.13%	55.78%	61.78%	75.48%	55.04%	58.26%	81.89%
MeNToS [18]	test	55.52%	49.98%	62.52%	52.38%	56.48%	77.13%	60.80%	64.97%	81.39%
SIA [23]	test	56.87%	51.36%	63.67%	55.83%	61.81%	75.52%	59.81%	63.63%	**83.38%**
SBT [26]	test	58.59%	53.20%	65.18%	55.92%	61.60%	75.93%	63.07%	68.54%	80.20%
Ours	test	**61.20%**	**55.57%**	**67.97%**	**59.43%**	**65.33%**	**78.16%**	**64.76%**	**69.47%**	83.03%

For "general", we follow the spirit of traditional reID methods [16, 17] that each object is cropped from the full image and resized to a fixed size to extract reID features individually. We do this for two reasons: 1) In most cases, general objects can be easily identified based on their colors, shapes and COCO categories. Thus, cropped object images mostly contain enough appearance information for reID. 2) Annotations of general objects in most MOTS datasets are much more sparse compared to common objects. Some of them (e.g. bags, hand-held tools) even have bounding box overlap with common objects. These may cause the reID learning of general objects misled by irrelevant contextual information dominated by common objects and backgrounds. To build this general reID model, we use Instance-Batch Normalization Network (IBN-Net50-A) [19] as its backbone and train it with a bag of tricks presented in [16]. Due to the relatively limited annotations of general objects in RobMOTS, we further collect more than 100,000 instance IDs from several video object detection datasets [22, 21] as the training data, which we believe would greatly enrich general object patterns including colors, shapes and categories. Based on these reID features, our general tracker uses DeepSORT [27] as its association strategy, where we also remove motion clues predicted by Kalman filter [1] similar to our common tracker.

2.2.2 Offline Re-Linking

To avoid accumulated errors in long-term tracking, an effective and necessary way is to set a maximum memory length N for our trackers. Specifically, if an ID has disappeared for more than N frames, it will be removed from the memory. However, this might cause broken tracklets when an ID reappears after being occluded for a long time, which occurs frequently in crowded scenes. To address this, we propose an Offline Re-Linking (ORL) strategy to fix these broken trajectories. For each video, we first compute average appearance features globally for all the IDs tracked online. If global appearance similarity between two IDs is greater than a threshold value (set to 0.8 in our experiments), we'll re-link them as the same ID. In practice, an ID can be matched with multiple other IDs with similarities above the threshold. In some circumstances, these IDs may have tracklet overlaps along time, which might cause the same ID be assigned to different objects in a frame. To avoid such ambiguity, we borrow the spirit from non-maximum suppression (NMS) algorithm [3, 8]. Firstly, we sort the matched IDs from large to small according to similarities. Then, if the matched IDs have time overlaps, we only assign the original ID to the most similar matched ID; otherwise we directly assign the original ID to all the matched IDs.

3. Experiments

3.1. Ablation Studies

As shown in Table 1, we conduct comprehensive ablation studies on the validation set of RobMOTS to investigate the importance of our major components. Specifically, we use Ensemble Full-Category (EFC) tracker provided with

original Cascade Mask R-CNN detections as a baseline, where the result is shown in the 1st row of Table 1. After progressively applying Massive Instance Augmentation (MIA), Patch-Based Mask Refinement (PBMR) and Offline Re-Linking (ORL) to the baseline, we can observe significant improvements on HOTA. In particular, MIA and PBMR greatly improve detection performance while ORL enables much better association performance. To further show the effectiveness of our ensemble tracker design, we replace the EFC tracker in our full method (the 6th row of Table 1) with single trackers whose reID models are trained in ways of common tracker and general tracker (the 4th row and the 5th rows of Table 1), respectively. The great improvements over the two variants with single trackers demonstrate the robustness of combining expert trackers from different category domains.

3.2. Comparison Results

We also compare our approach with other top methods [26, 23, 18, 13] in the RobMOTS challenge on validation and test sets. As can be seen from Table 2, our RobTrack baseline significantly outperforms all of the other methods for not only detection accuracy (DetA) but also association accuracy (AssA), indicating its effectiveness in both detection and association for robust tracking.

4. Conclusion

In this paper, we propose RobTrack, a robust tracker baseline for real-world Multi-Object Tracking and segmentation (MOTS). We divide the overall MOTS task into detection and association, and optimize both of them to achieve robust tracking. Extensive experiments demonstrate our effectiveness and superiority to other approaches in robust real-world tracking.

References

[1] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 2, 3

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2

[3] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 3

[4] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *European conference on computer vision*, pages 436–454. Springer, 2020. 2

[5] Achal Dave, Pavel Tokmakov, and Deva Ramanan. Towards segmenting anything that moves. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[6] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. *arXiv preprint arXiv:2012.07177*, 2020. 2

[7] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7036–7045, 2019. 2

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 3

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2

[12] Jonathon Luiten, Tobias Fischer, and Bastian Leibe. Track to reconstruct and reconstruct to track. *IEEE Robotics and Automation Letters*, 5(2):1803–1810, 2020. 2

[13] Jonathon Luiten, Arne Hoffhues, Blin Beqa, Paul Voigtlaender, István Sárándi, Patrick Dendorfer, Aljosa Osep, Achal Dave, Tarasha Khurana, Tobias Fischer, Xia Li, Yuchen Fan, Pavel Tokmakov, Song Bai, Yang Linjie, Federico Perazzi, Ning Xu, Alex Bewley, Jack Valmadre, Sergi Caelles, Jordi Pont-Tuset, Xinggang Wang, Andreas Geiger, Fisher Yu, Deva Ramanan, Laura Leal-Taixé, and Bastian Leibe. Robmots: A benchmark and simple baselines for robust multi-object tracking and segmentation. In *CVPR RVSU Workshop*, 2021. 1, 3, 4

[14] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2):548–578, 2021. 1

[15] Jonathon Luiten, Idil Esen Zulfikar, and Bastian Leibe. Unovost: Unsupervised offline video object segmentation and tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2000–2009, 2020. 2

[16] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[17] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*, pages 1–1, 2019. 3

[18] Mehdi Miah, Guillaume-Alexandre Bilodeau, and Nicolas Saunier. Mentos : Tracklets association with a space-time memory network. In *CVPR RVSU Workshop*, 2021. 3, 4

[19] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 3

[20] Jiangmiao Pang, Linlu Qiu, H Chen, Q Li, T Darrell, and F Yu. Quasi-dense similarity learning for multiple object tracking. *preprint*, 2020. 2

[21] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5296–5305, 2017. 3

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3

[23] Jeongwon Ryu and Kwangjin Yoon. Sia: Simple re-identification association for robust multi-object tracking and segmentation. In *CVPR RVSU Workshop*, 2021. 3, 4

[24] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 2

[25] Chufeng Tang, Hang Chen, Xiao Li, Jianmin Li, Zhaoxiang Zhang, and Xiaolin Hu. Look closer to segment better: Boundary patch refinement for instance segmentation. *arXiv preprint arXiv:2104.05239*, 2021. 2

[26] Jiasheng Tang, Fei Du, Weihua Chen, Hao Luo, Fan Wang, and Hao Li. Sbt: A simple baseline with cascade association for robust multi-objects tracking. In *CVPR RVSU Workshop*, 2021. 3, 4

[27] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 3